The Role of Ethical Design in Building Trust in Machine Learning

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Akshay Choksi

Spring, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

Ethical Inquiry into Machine Learning Algorithms

In the ever-expanding realm of technology, the integration of machine learning algorithms has become ubiquitous, promising unprecedented advancements in efficiency and automation. However, as these algorithms become a part of various sectors of society, questions about their ethical implications have taken center stage. From biased decision-making to nontransparent processes, there are numerous concerns regarding the design and deployment of machine learning systems (Rader & Gray, 2015). This study delves into the intersection of technology and ethics, aiming to explain the complexities surrounding the integration of ethical considerations into the design of machine learning algorithms. Using Pinch and Bijker's Social Construction of Technology (SCOT) framework, the research seeks to understand how different societal stakeholders shape and interpret the foundational dimensions of machine learning, ultimately guiding the formulation of guidelines for ethically sound technological design practices. The central question guiding this investigation is: How can the ethical considerations surrounding machine learning algorithms be integrated into their technological design to increase stakeholder trust?

Investigation of Ethical Design

The methodology in this research involves a multi-faceted approach to understanding ethical considerations in the technological design of machine learning tools. Case studies and a historiography of relevant literature are utilized as primary methods to gather data and insights. Case studies provide in-depth examinations of real-world scenarios where ethical design practices have been implemented or faced challenges. Additionally, a historiography of scholarly works and industry reports guides the exploration of key concepts, trends, and debates in the field of AI ethics. Keywords such as "ethical design," "machine learning," "stakeholder trust," "transparency," and "accountability" guide the research inquiry and data collection process. To

analyze the results with the SCOT framework, this paper is divided into sections by relevant social groups and further combined with a wider context in society that will be used to answer the research question.

Foundations of Machine Learning Design

To establish a comprehensive understanding of the research context, it is essential to delve into the foundational concepts of machine learning and its inescapable influence. Machine learning, a subset of artificial intelligence (AI), entails the development of algorithms and statistical models that enable computer systems to perform tasks without explicit programming instructions. These complex algorithms learn from data, identifying patterns and making predictions or decisions based on their analysis. The applications of machine learning span diverse domains, from recommendation systems in e-commerce to medical diagnosis and autonomous driving.

Within the realm of machine learning, the concept of ethical design has emerged as a critical consideration. In the context of machine learning algorithms, ethical design involves mitigating biases, ensuring transparency in decision-making processes, safeguarding user privacy, and fostering accountability for the outcomes generated by these algorithms (Uddin, 2023). The proliferation of machine learning technologies has prompted heightened scrutiny regarding their ethical implications. Instances of algorithmic bias, where machine learning systems perpetuate or exacerbate discriminatory outcomes, have raised concerns about fairness and equity (Chen, 2023). Additionally, the opacity of some machine learning models, often referred to as "black boxes," poses challenges to understanding how decisions are made, limiting accountability and trust (Rudin & Radin, 2019). Furthermore, the vast quantities of data

processed by machine learning algorithms raise questions about data privacy and security, particularly in contexts involving sensitive personal information.

In response to these ethical challenges, scholars, policymakers, and industry stakeholders have called for greater attention to ethical foresight in the design and deployment of machine learning systems (Gerke et al., 2020). Regulatory frameworks, such as the General Data Protection Regulation (GDPR) in the European Union, aim to protect individual rights and promote ethical data practices. Moreover, interdisciplinary research efforts have emerged to develop guidelines, frameworks, and tools for ethical design in machine learning, emphasizing the importance of collaboration and stakeholder engagement in addressing ethical concerns.

This study seeks to contribute to the ongoing discourse on ethical technology development. Through an exploration of stakeholder perspectives and the application of the Social Construction of Technology (SCOT) framework, the research aims to offer insights into how ethical considerations can be integrated into the technological design of machine learning algorithms, fostering trust and accountability in AI-driven systems.

Social Construction of Technology and Machine Learning

Ethical considerations surrounding machine learning algorithms align closely with Science and Technology Studies (STS) principles, as they highlight the complex interactions between technology and society, particularly regarding power relations, values, and societal impacts (Jasanoff & Kim, 2015). By examining the ethical dimensions of machine learning through an STS lens, this study aims to highlight the social construction of technology, emphasizing the importance of societal values and norms in shaping technological advancements.

SCOT argues that technology is not inherently determined but rather develops through the interactions between various social actors and the broader socio-cultural context. Scholars within the SCOT framework emphasize the socially applicable nature of technological development, highlighting how different social groups negotiate and interpret technologies based on their interests, values, and power dynamics (Pinch & Bijker, 1984). The SCOT framework has been widely used in STS to analyze the emergence and evolution of various technologies, including the internet, mobile phones, and renewable energy systems. A literature review analyzed negative outcomes of AI technology through job displacement (Tiwari, 2023). Tiwari found that a more in-depth approach "could also focus on the impact of AI and ML on different marginalized groups," suggesting that SCOT has value to the research topic at-hand (Tiwari, 2023). By applying SCOT to the study of ethical design in machine learning algorithms, this research seeks to understand how societal values, ethical norms, and power relations influence the development and deployment of AI technologies.

Discussion of the Considerations Required

Ethical design involves a varied approach that requires input and considerations from pertinent stakeholders to ensure the ethical development of machine learning algorithms. While each relevant social group has their intended performance and concerns with production and consumption, each has a part to play in its design process - from start to finish. Interpretative flexibility, therefore, has to be analyzed in regards to each group to understand how to achieve stabilization in the context of ethical design. Also, it should be compared to realize where the considerable faults lie. Moreover, from a SCOT perspective, the societal impact must be weighed in the “wider context” as this is the stage of which “artifact development takes place” (Klein & Kleinman, 2002). Exploring the multifaceted landscape of algorithm designers, software users,

and technology policymakers, this discussion delves into how their diverse perspectives and priorities contribute to the ethical design process in machine learning algorithms.

Algorithm Designers

The developers of machine learning software have been around for decades now, yet the design intuition and structures have seen little progress in terms of ethical foresight. While new ideas are presented "daily," the new aspect is how they are being "applied at scale," which government regulators have struggled to keep pace with (Bell, 2020). Major technology companies have continued to dominate the media for privacy violations such as Google and Facebook. A developer's work involves not just the technical aspects of creating efficient and effective algorithms but also understanding the broader implications of how these algorithms are used.

In the context of machine learning algorithm design, interpretive flexibility allows for a broad spectrum of design motivations and outcomes, influenced by varying priorities. One important example is seen in the approach of companies like Facebook, where the design of algorithms is often interpreted through the lens of profitability and business models. Lauer (2021) critically examines Facebook's "failures," suggesting that they are not merely accidental but are deeply embedded within the "company's profit-oriented business model." This perspective demonstrates a social group within the algorithm design community that prioritizes financial success over ethical considerations. The profit-driven interpretation of algorithm design is defined by a focus on metrics such as user engagement, ad targeting efficiency, and data monetization. In this model, ethical design is secondary to the primary goal of maximizing revenue. The controversy surrounding Facebook's handling of user data and its impact on

privacy violations is a testament to the consequences of a design philosophy that places profit above ethical considerations.

However, there are other interpretations within this group that require a different discussion. In pediatric settings, the stakes of ethical design are particularly high, given the vulnerability of the patient population (Nsier, 2023). The study highlights the need for algorithm designers to engage with a wide range of stakeholders - including clinicians, patients, and legal guardians - to ensure that the algorithms are designed with the highest consideration for ethical principles. Similarly, the EvoMol study on a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation highlights the importance of interpretive flexibility in the realm of drug discovery (Cheminform, 2020). The development of unbiased algorithms in this context requires a deep understanding of the ethical implications associated with drug design, including considerations around equity in access to the resulting medications and the potential for unintended consequences. By incorporating ethical considerations into the design process, algorithm designers work towards solutions that are not only innovative but also equitable and socially responsible. Therefore, these developers benefit from being proactive in their ethical design serving as a counterexample to a company such as Facebook.

One notable example of ethical design becoming a part of stabilization is the development of the AI system, DALL-E, by OpenAI, which generates images from textual descriptions. OpenAI implemented a series of ethical design considerations to mitigate potential misuse, such as filtering out explicit content and ensuring diversity in the generated images (Roumeliotis & Tselikas, 2023). This approach demonstrates how proactive engagement with ethical considerations can lead to the development of AI technologies that are both innovative and responsible. In the healthcare sector, IBM Watson for Oncology assists clinicians in

identifying personalized cancer treatment options. However, its effectiveness and ethical implications have been scrutinized, particularly regarding transparency and the evidence base for its recommendations. In response, IBM has engaged in ongoing efforts to improve the system's transparency and reliability, working closely with healthcare professionals to ensure that the technology aligns with both ethical standards and clinical needs (Thomas et al., 2023). This case illustrates the dynamic process of stabilization, where continuous feedback is essential for integrating ethical considerations into technological design.

Software Users

As the end consumers of machine learning technologies, a user's experiences significantly influence the acceptance of these systems. Users are aware of the potential for algorithms to perpetuate biases and the challenges posed by opaque, "black box" models that offer little insight into their decision-making processes. This awareness has led to a demand for greater transparency, accountability, and fairness in machine learning technologies, which in turn shapes the ethical design considerations that developers must address.

Some users may prioritize the accuracy and efficiency of machine learning systems, willing to trade off some degree of transparency for better performance. Others may be more concerned with understanding how decisions are made, especially in high-stakes scenarios such as medical diagnoses or criminal justice. This diversity in user expectations and values, therefore, requires a design approach that is flexible and responsive to a wide range of ethical concerns. Rader and Gray (2015) discuss how users' trust in algorithms can be influenced by their understanding of the system's functionality and their own perceived fairness of its outcomes. Users who are provided with clear explanations of how an algorithm works and how their data is used are more likely to trust and then accept the technology. On the other hand, a

lack of transparency can lead to resistance, especially if users suspect that the system may be biased or affecting their privacy.

Achieving stabilization and closure in the context of software users involves the development of machine learning algorithms that consistently meet ethical expectations across diverse user groups. This requires a coordinated effort to address interpretive flexibility by incorporating user feedback into the design process and providing simple explanations of how algorithms function. One approach to achieving stabilization is through the implementation of explainable AI (XAI) systems, which aim to make the decision-making processes of machine learning algorithms more transparent and understandable to users (Arrieta et al., 2020). By providing users with insights into how algorithms generate their outputs, XAI can help showcase the technology and build trust. For example, Rudin and Radin (2019) emphasize the importance of explainability in fostering user trust and enabling stakeholders to assess the fairness and accountability of machine learning systems. Furthermore, addressing bias in algorithms is crucial for achieving ethical stabilization. Chen (2023) highlights the need for machine learning systems to be designed with mechanisms that detect and mitigate biases, ensuring that the technology operates fairly and equitably. This involves not only technical solutions but also engaging with users to understand their perspectives on what exactly constitutes bias and fairness.

Technology Policymakers

Lawmakers play a critical role in shaping the ethical landscape of machine learning technologies. They are responsible for creating regulatory frameworks that protect individual rights and promote ethical data practices. Policymakers must balance the potential benefits of technology, such as economic growth and improved public services, with the risks, including privacy concerns and potential job displacement.

For instance, the Congressional Research Service report on science and technology policymaking highlights the democratic nature of the decision-making process, where a wide array of organizations and individuals contribute diverse ideas and opinions. This diversity leads to different interpretations of what constitutes the most pressing issues and the best strategies for addressing them (Stine, 2009). Policymakers must navigate a complex landscape of competing interests and values, which results in a range of policy outcomes. The interpretive flexibility of technology policy is also discussed in the context of the UK's National Health Service Care Records Service. The study by Papazafeiropoulou et al. reveals contrasting views on the system's implementation at local and national levels, demonstrating how interpretive flexibility manifests through the diverse perceptions of stakeholders involved in the adoption of technology. It explains that the success was largely due to patients needing to be clearly informed of the benefits of the technology and how it can legitimately benefit them.

Stabilization in technology policymaking occurs when consensus is reached, and policies are enacted to guide the development and use of technology. This process often involves negotiations and compromises among various stakeholders, including government agencies, industry leaders, and the public. International collaboration plays a crucial role in achieving stabilization in machine learning policy. Given the global nature of machine learning development, international agreements and frameworks can help harmonize regulatory approaches and ensure that ethical considerations are consistently addressed worldwide. This collaborative effort can also facilitate the sharing of best practices and lessons learned among countries, contributing to a more cohesive and effective global response to the challenges posed by machine learning (Dickow & Jacob, 2018). The General Data Protection Regulation (GDPR) serves as a critical reference point for ethical stabilization in machine learning policy. It provides

a framework for data protection, granting individuals greater control over their personal data and imposing strict requirements on data processors. By mandating transparency, consent, and accountability, the GDPR has set a global standard for privacy and data protection that machine learning developers and policymakers must consider in the design and regulation of systems.

Forming an Ethical Design for Society

The "wider context" as described in SCOT, includes the broader sociocultural and political environment in which the development and implementation of machine learning technologies occur (Klein & Kleinman, 2002). This context plays a crucial role in shaping the interactions among the three primary social groups involved in the ethical design of machine learning tools: algorithm designers, software users, and technology policymakers. It influences their perspectives, actions, and the dynamics of their relationships, ultimately affecting the stabilization of ethical considerations in machine learning technologies.

For algorithm designers, the wider context includes the prevailing industry norms, competitive pressures, and the broader societal expectations regarding technology and ethics. Companies like Facebook and Google operate within a highly competitive tech industry where innovation, user engagement, and profitability are often prioritized. This environment can influence algorithm designers to prioritize business objectives over ethical considerations, as seen in Facebook's profit-oriented algorithm design. However, the wider context also includes a growing public and academic discourse on the ethical implications of AI, exemplified by studies on the use of AI in pediatric medical settings and drug discovery. This discourse is gradually shaping a new normative framework that emphasizes the importance of ethical design, pushing algorithm designers to integrate ethical considerations proactively, as demonstrated by OpenAI's development of DALL-E.

On the other hand, software users are influenced by their sociocultural backgrounds, personal experiences, and the general public's sentiment towards technology. The wider context for users includes the increasing awareness and concern over data privacy, algorithmic bias, and the ethical use of AI. This is fueled by high-profile incidents of data misuse and biased algorithms. Users' trust in algorithms can be significantly affected by their understanding of how these systems work and the perceived fairness of their outcomes. The demand for explainable AI (XAI) systems and mechanisms to detect and mitigate biases reflects the wider societal push for more ethical and understandable AI technologies.

Technology policymakers operate within a complex landscape of legal, political, and societal pressures. The wider context for policymakers includes the challenge of keeping pace with rapid technological advancements while addressing public concerns about privacy, security, and ethical implications. International collaboration and agreements, such as the GDPR, highlight how the wider context influences policy development by setting global standards for privacy and data protection. The process of achieving stabilization in AI policy is therefore embedded in the wider sociocultural and political environment. It requires a nuanced understanding of the complex relationship between technology, society, and ethics.

Future of Design Research

While this analysis found valuable sources and case studies to depict an ethical approach to designing and maintaining machine learning systems, there are a few limitations. Firstly, the machine learning community is rapidly advancing with new case studies and policies being introduced almost constantly. Therefore, it is challenging to apply certain designs when they could be dated or replaced. Moreover, this research focused on three relevant social groups, when there are many more factors and groups of people that have influence over the ethical

design of machine learning technology. Further, the generalization of groups, while simplifying the application of the SCOT framework, limits the distinction between different types of users of machine learning technology which could be valuable research.

There is a need for in-depth empirical studies that examine the actual implementation of ethical design principles in machine learning systems across different industries and contexts. This would involve collaborating closely with industry practitioners to understand the challenges, best practices, and impact of ethical design on algorithm performance and user experiences. Furthermore, interdisciplinary collaborations between ethicists, computer scientists, and policymakers are essential for developing holistic frameworks and guidelines for ethical AI development and deployment. Lastly, studies that track the evolution of ethical considerations in machine learning algorithms over time can provide valuable insights into trends, challenges, and areas for improvement in ethical design practices.

Final Takeaways of Research

This research underscores the critical importance of integrating ethical considerations into the technological design of machine learning tools to bolster stakeholder trust. Through delving into the perspectives and roles of algorithm designers, software users, and technology policymakers, the research found there are intricate challenges and nuances associated with ethical design practices in AI. The key takeaway is the need for proactive engagement with ethical principles, transparent decision-making processes, accountability mechanisms, and a user-centered approach to algorithm design. These foundational aspects not only contribute to enhancing trust and acceptance of AI technologies but also pave the way for responsible innovation and sustainable deployment of machine learning tools. This research not only enriches the academic discourse on ethics in machine learning but also provides actionable

insights and recommendations for industry practitioners, policymakers, and researchers navigating the ethical complexities of modern systems.

Works Cited:

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58, 82-115.

Bell, J. (2020). Machine Learning: Hands-on for Developers and Technical Professionals. J*ohn Wiley & Sons, Inc*.

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1). https://doi.org/10.1057/s41599-023-02079-x

Dickow, M., & Jacob, D. (2018). The global debate on the future of artificial intelligence: the need for international regulation and opportunities for German foreign policy.

Gerke, S., Minssen, T., & Cohen, G. (2020b). Ethical and Legal Challenges of Artificial Intelligence-driven Healthcare. *Artificial Intelligence in Healthcare*, 295–336. https://doi.org/10.1016/b978-0-12-818438-7.00012-5

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., & Farkash, A. (2021). Data Minimization for GDPR compliance in machine learning models. *AI and Ethics*, *2*(3), 477–491. https://doi.org/10.1007/s43681-021-00095-8

Jasanoff, S., & Kim, S.-H. (Eds.). (2015). Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. *University of Chicago Press*.

Klein, H. K., & Kleinman, D. L. (2002). The social construction of Technology: Structural Considerations. *Science, Technology, Human Values*, *27*(1), 28–52. https://doi.org/10.1177/016224390202700102

Lauer, D. (2021). Facebook's ethical failures are not accidental; they are part of the business model. *AI Ethics* 1, 395–403. https://doi.org/10.1007/s43681-021-00068-x

Mourby, M., Ó Cathaoir, K., & Collin, C. B. (2021, November 1). Transparency of machine-learning in healthcare: The GDPR & European health law. *Computer Law & Security Review: The International Journal of Technology Law and Practice*, 43, 1 - 14. https://doi.org/10.1016/j.clsr.2021.105611

Pinch, T. J., & Bijker, W. E. (1984, August 1). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science,* 14(3), 399 - 441.

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the facebook news feed. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2702123.2702174

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Tiwari, R. (2023). Ethical and societal implications of AI and machine learning. *International Journal of Scientific Research in Engineering and Management*, 07(01), 20 - 27. https://doi.org/10.55041/ijsrem17519

Uddin, A. (2023) The Era of AI: Upholding Ethical Leadership. *Open Journal of Leadership*, 12, 400-417. doi: 10.4236/ojl.2023.124019.

Taeihagh, A. (2021). Governance of Artificial Intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Thomas, D. M., Kleinberg, S., Brown, A. W., Crow, M., Bastian, N. D., Reisweber, N., Lasater,

R., Kendall, T., Shafto, P., Blaine, R., Smith, S., Ruiz, D., Morrell, C., & Clark, N. (2022). Machine learning modeling practices to support the principles of AI and ethics in nutrition research. *Nutrition & Diabetes*, *12*(1). https://doi.org/10.1038/s41387-022-00226-y

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key Problems and Solutions. *AI & SOCIETY*, *37*(1), 215–230. https://doi.org/10.1007/s00146-021-01154-8

Roumeliotis, K., & Tselikas, N. (2023, May 26). *Chatgpt and open-AI models: A preliminary review*. MDPI. https://www.mdpi.com/1999-5903/15/6/192

Stine, D. (2009, May 27). Science and Technology Policymaking: A Primer. *Congressional Research Service*.